CNApp: a web-based tool for integrative analysis of genomic copy number alterations in cancer

Sebastià Franch-Expósito[1*], Laia Bassaganyas[2*], Maria Vila-Casadesús[3], Eva Hernández-Illán[1], Roger Esteban-Fabró[2], Marcos Díaz-Gay[1], Juan José Lozano[3], Antoni Castells[1], Josep M. Llovet[2,4,5], Sergi Castellví-Bel[1#], Jordi Camps[1,6#]

1 Gastrointestinal and Pancreatic Oncology Team, Institut D'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic de Barcelona, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Universitat de Barcelona, Barcelona, Catalonia, Spain

2 Liver Cancer Translational Research Group, Liver Unit, Institut D'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), Universitat de Barcelona, Barcelona, Catalonia, Spain

3 Bioinformatics Unit, CIBEREHD, Barcelona, Catalonia, Spain

4 Mount Sinai Liver Cancer Program, Division of Liver Diseases, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, USA

5 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

6 Unitat de Biologia Cel·lular i Genètica Mèdica, Departament de Biologia Cel·lular, Fisiologia i Immunologia, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra, Catalonia, Spain

* equal contribution; # corresponding author

Running title: CNApp: copy number integrative analysis

Keywords: Copy number alterations, hepatocellular carcinoma, colorectal cancer, pan-cancer, machine learning

ABSTRACT

Copy number alterations (CNAs) are a hallmark of cancer. Large-scale cancer genomic studies have already established the CNA landscape of most human tumor types and some CNAs are recognized as cancer-driver events. However, their precise role in tumorigenesis as well as their clinical and therapeutic relevance remain undefined, thus computational and statistical approaches are required for the biological interpretation of these data. Here, we describe CNApp, a user-friendly web tool that offers sample- and cohort-level association analyses, allowing a comprehensive and integrative exploration of CNAs with clinical and molecular variables. CNApp generates genome-wide profiles, calculates CNA levels by computing broad, focal and global CNA scores, and uses machine learning-based predictions to classify samples by using segmented data from either microarrays or next-generation sequencing. In the present study, using copy number data of well-annotated 10,635 genomes from The Cancer Genome Atlas spanning 33 cancer subtypes, we showed that patterns of CNAs classified tumor subtypes according to their tissue-of-origin and that broad and focal CNA scores correlated positively in those samples with low levels of chromosome and arm-level events. Moreover, CNApp allowed the description of recurrent CNAs in hepatocellular carcinoma further confirming previous results identified using other methods. Finally, we established machine learning-based models to predict colon cancer molecular subtypes and microsatellite instability based on broad and focal CNA scores and specific genomic imbalances. In summary, CNApp facilitates data-driven research and provides a unique framework to comprehensively assess CNAs and perform integrative analyses that enable the identification of relevant functional implications.

INTRODUCTION

The presence of somatic copy number alterations (CNAs) is a ubiquitous feature in cancer. Indeed, the distribution of such CNAs is sufficiently tissue-specific to distinguish and enable the classification of tumor entities (Ried et al. 2012; Taylor et al. 2018a), and may allow identifying groups of tumors responsive to particular therapies(Cairncross et al. 2013; Davoli et al. 2017). Moreover, high levels of CNAs, which result from aneuploidy and chromosome instability, are generally associated with high-grade tumors and poor prognosis (Sansregret et al. 2018). Two main subtypes of CNAs can be discerned: broad CNAs, which are defined as whole-chromosome and chromosomal arm-level alterations, and focal CNAs, which are alterations of limited size ranging from part of a chromosome-arm to few kilobases (Krijgsman et al. 2014; Zack et al. 2013). Recently, it has been uncovered that while focal events mainly correlate with cell cycle and proliferation markers, broad aberrations are mainly associated with immune evasion markers (Taylor et al. 2018b; Davoli et al. 2017; Buccitelli et al. 2017). Nevertheless, the precise role of CNAs in tumor initiation and progression, as well as their clinical relevance and therapeutic implications remain still poorly understood.

Characterization and interpretation of CNAs is time-consuming and very often requires complex integrative analyses with clinical and molecular information. Moreover, visualization of complex data is usually essential to discriminate key results. Well-established CNA algorithms, such as the gold-standard circular binary segmentation, determine the genomic boundaries of copy number gains and losses based on signal intensities or read depth obtained from array comparative genomic hybridization and SNP-array or next-generation sequencing data, respectively (Olshen et al. 2004). Variability within gain or loss levels can be addressed with the algorithm CGHcall, which enables the identification of single copy number changes (van de Wie et al. 2007). In order to overcome the complex nature of tumor samples (Stratton et al. 2009), more recent segmentation methods improved the accuracy to identify copy number segments either by considering the B allele frequency (BAF), such as ExomeCNV (Sathirapongsasuti et al. 2011), Control-FREEC (Boeva et al. 2012) and SAAS-CNV (Zhang and Hao 2015), or through adjusting by sample purity and ploidy estimates, such as GAP (Popova et al. 2009), ASCAT (Van Loo et al. 2010) and ABSOLUTE (Carter et al. 2012). However, the state-of-the-art computational approach for CNA analysis is GISTIC2.0 (Mermel et al. 2011), which is a gene-centered probabilistic method that enables to define the boundaries of recurrent putative driver CNAs in large cohorts (Beroukhim et al. 2010). Nevertheless, despite ongoing progress on identifying CNAs, to our knowledge there is no bioinformatic tool readily available for integrative analyses to unveil the biological interpretation of these CNAs.

To address this issue, we developed CNApp, the first open-source application to comprehensively analyze and integrate CNA profiles with molecular and clinical variables. CNApp was built in Shiny R package (Chang et al. 2018) and provides the user with high-quality interactive plots and statistical correlations between CNAs and annotated variables in a fast and easy-to-explore interface. In particular, CNApp uses genomic segmented data to quantify CNA levels based on broad and focal genomic alterations, assess differentially altered genomic regions, and perform machine learning-based predictions to classify tumor samples. A dataset including 160 colon cancer samples with clinical annotation is loaded for demonstration purposes. To exemplify the applicability and performance of CNApp, we used publicly available segmented data from The Cancer Genome Atlas (TCGA) to (i) measure the burden of global, broad, focal CNAs as well as generate CNA profiles in a pan-cancer dataset spanning 33 cancer subtypes, (ii) identify cohort-based recurrent CNAs in hepatocellular carcinoma and compare it with previously reported data using different methods, and (iii) assess predicting models for colon cancer molecular subtype and microsatellite instability status classification based on CNA scores and specific genomic imbalances. CNApp is hosted at http://bioinfo.ciberehd.org/CNApp and the source code is freely available at GitHub (https://github.com/ait5/CNApp).

RESULTS

Implementation and basic usage

Functions of CNApp comprise three main sections: 1- *Re-Seg & Score: re-segmentation, CNA scores computation and variable association*, 2- *Region profile: genome-wide CNA profiling*, and 3- *Classifier model: machine learning classification model predictions*, (Figure 1). Each of these sections and their key functions are described below. The input file consists of a data frame with copy number segments provided by any segmentation algorithm. Mandatory fields and column headers are sample name (*ID*), chromosome (*chr*), start (*loc.start*) and end (*loc.end*) genomic positions, and the log2 ratio of the copy number amplitude (*seg.mean*) for each segment. If available, it is recommended to include sample purity (*purity*) and BAF values (*BAF*), which can improve the accuracy of CNA calls and will provide information of copy number neutral loss-of-heterozygosity (CN-LOH) events. Annotation of variables can be included in the input file (tagged in every segment from each sample) or by loading an additional file specifying new variables to every sample.

Section 1. Re-Seg & Score: re-segmentation, CNA scores computation and variable association

First, CNApp applies a re-segmentation approach aiming at correcting potential background noise and amplitude divergence due to technical variability. Default re-segmentation settings include *minimum segment length* (100 Kbp), *maximum distance between segments* (1 Mbp), *maximum amplitude (seg.mean) deviation between segments*

(0.16), *minimum amplitude (seg.mean) deviation from segment to zero* (0.16), and *maximum BAF deviation between segments* (0.1). These parameters can be customized by the user to better adjust the re-segmentation and CNA calling for each particular dataset. Re-segmented data are then used to calculate the focal (FCS), broad (BCS) and global (GCS) CNA scores, which provide three different quantification of CNA levels for each sample. To compute these scores, CNApp classifies and weights CNAs based on amplitude and length. A weight is given to each segment according to its *seg.mean* value and by applying low-, medium- and high-level copy number amplitude thresholds. By considering the relative length of each segment to the whole-chromosome or chromosome arm, segments are tagged as *chromosomal* -by default, 90% or more of the chromosome affected-, as *arm-level* -50% or more of the chromosome arm affected-, or as *focal* -less than 50% of the chromosome arm affected. Percentages for relative lengths are also customizable. For each sample, BCS is computed by considering chromosome and arm-level segment weights according to the amplitude value. Likewise, calculation of FCS takes into account weighted focal CNAs and the amplitude and length of the segment. Finally, GCS is computed by considering the sum of normalized FCS and BCS values, providing an overall assessment of the CNA burden for each sample. To assess the reliability of CNA scores, we compared each score with the corresponding fraction of altered genome using a TCGA pan-cancer set of 10,635 samples. Both FCS (values ranging from 5 to 2,466) and BCS (ranging from 0 to 44) highly correlated with the fraction of altered genome by focal and broad copy number changes, respectively (Spearman's rank correlation for BCS = 0.957 and for FCS = 0.938) (Supplemental_Fig_S1 A and B). As expected, GCS (values ranged from -1.93 to 12.60) highly correlated with the fraction of altered genome affected by both focal and broad CNAs (Spearman’s rank correlation for GCS = 0.963) (Supplemental_Fig_S1C).

Additionally, parametric and non-parametric statistical tests are used to establish associations between CNA scores and annotated variables from the input file.

Section 2. Region profile: genome-wide CNA profiling

This section utilizes re-segmented data obtained from section 1 or uploaded segmented data without re-segmentation to generate genomic region profiling and sample-to-sample correlations. To conduct this, re-segmented data are transformed into genome region profiles according to a user-selected genomic window (i.e., chromosome arms, half-arms, cytobands, sub-cytobands or 40-1 Mbp windows). All segments, or either only broad or only focal can be selected for this analysis. Length-relative means are computed for each window by considering amplitude values from those segments included in each specific window. Default thresholds for low-level copy number gains and losses (i.e., |0.2|) are used as cutoffs to classify genome regions and to calculate their frequencies in this section. Genome-region profiles are presented in genome-wide heatmaps to visualize general copy number patterns. Up to six annotation tracks can be added and plotted simultaneously allowing visual comparison and correlation between CNA profiles and different variables, including the CNA scores obtained in section 1. Generation of hierarchical clusters by samples and regions is optional. CNA frequency summaries by genomic region and by sample are represented as stacked bar plots.

Importantly, assessing differentially altered regions between sample groups might contribute to discover genomic regions associated with annotated variables and thus unveil the biological significance of specific CNAs. To do so, CNApp interrogates descriptive regions associated with any sample-specific annotation variable provided in the input file. Student's t-test or Fisher's test are applied when considering CNAs as continuous alterations (*seg.mean* values) or as categorical events (presence of gains and losses), respectively. Default statistical significance is set to *P*-value lower than 0.1.

However, p-value thresholds can be defined by the user and adjusted *P*-value is optional. A heatmap plot allows the visualization and interpretation of which genome regions are able to discriminate between sample groups. By selecting a region of interest, box plots and stacked bar plots are generated comparing *seg.mean* values and alteration counts in Student's t-test and Fisher's test tabs, respectively. Additionally, genes comprised in the selected region are indicated.

3. Classifier model: Machine learning classification model predictions

This section allows the user to generate machine learning-based classifier models by choosing a variable to define sample groups and one or multiple classifier variables. To do so, CNApp incorporates the *randomForest* R package (Liaw and Wiener 2002). The model construction is performed 50-times and bootstrap set is changed in each iteration. By default, only annotation variables from the input file are loaded to work either by group defining or by classifier variables. If *Re-Seg & Score* and/or *Region profile* sections have been previously completed, the user can upload data from these sections (i.e., CNA scores and genomic regions). Predictions for the model performance are generated and the global accuracy is computed along with sensitivity and specificity values by group. Classifier models can be useful to point out candidate clinical or molecular variables to classify sample subgroups. A summary of the data distribution and plots for real and model-predicted groups are visualized. A table with prediction rates throughout the 50-times iteration model and real tags by sample is displayed and can be downloaded.

Genomic characterization of cancer subtypes

First, we evaluated the capacity of CNApp to analyze and classify cancer subtypes according to distinct patterns of CNA scores, and assess whether CNApp was able to reproduce the distribution of cancer subtypes based on specific CNA profiles. To do so, level 3 publicly available Affymetrix SNP 6.0 array data from 10,635 tumor samples spanning 33 cancer types from TCGA pan-cancer database were used. We applied *Re-Seg & Score* and *Region profile* using default parameters to obtain re-segmented data, CNA scores, and cancer-specific CNA profiles. Correlations between CNA scores were assessed by computing Spearman's rank test, obtaining values of 0.59 between BCS and FCS, 0.90 between BCS and GCS, and 0.85 between FCS and GCS. In addition, we further assessed the correlation between BCS and FCS for each individual BCS value. While tumors with low BCS displayed a positive correlation between broad and focal alterations, tumors did not maintain such correlation in higher BCS values (Supplemental_Fig_S2A). BCS, FCS and GCS distributions across cancer subtypes supported the existence of distinct CNA levels between tumors from different origin (Figure 2A). While cancer subtypes such as acute myeloid leukemia (LAML), thyroid carcinoma (THCA) or thymoma (THYM) showed low levels of broad and focal events (GCS median values of -1.67 for LAML, -1.68 for THCA, and -1.52 for THYM), uterine carcinosarcoma (UCS), ovarian cancer (OV) and lung squamous cell carcinoma (LUSC) displayed high levels of both types of genomic imbalances (GCS median values of 2.55, 2.44, and 0.97 for UCS, OV, and LUSC, respectively). Some cancer subtypes displayed a preference for either broad or focal copy number alterations. For example, kidney chromophobe (KICH) tumors showed the highest levels of broad events (median BCS value of 27); however, they were amongst those subtypes with less focal CNAs (median FCS value of 49). In contrast, breast cancer (BRCA) samples displayed high values for FCS (median FCS value of 150), while BCS values were intermediate (median BCS value of 7).

Subsequent analysis aimed at generating genome-wide patterns for each cancer subtype based on chromosome-arm genomic windows and the overall corresponding frequencies (Figure 2B). We found that chromosome arms altered in more than 25% across all samples were 1q, 7p, 7q, 8q and 20q for copy number gains, and 8p and 17p for copy number losses. Conversely, chromosome arms affected by CNAs in less than 10% of all cancer subtypes included 2q and 19p (Figure 2C). By using a subset of 20 out of the 33 cancer types for which tumor type information was available, we asked CNApp to compute the average arm-region for each cancer type to assess if they clustered according to their CNA profile (Supplemental_Fig_S2B). Our analysis showed that correlation profiles resulting from Pearson's test were hierarchically clustered according to their tumor type (Figure 2D). Gastrointestinal (colon, rectum, stomach and pancreatic), gynecological (ovarian and uterine) and squamous (cervical, head and neck, and lung) cancers clustered together based on specific CNA profiles for each group (Figure S2B). These results strongly correlated with previously reported findings (Taylor et al. 2018b; Hoadley et al. 2018).

Identification of recurrent CNAs in liver hepatocellular carcinoma

Next, we attempted to test the ability of CNApp to identify recurrent broad and focal CNAs in a large cohort of samples. For that reason, we chose to perform CNA analysis of 370 samples from TCGA corresponding to the Liver Hepatocellular Carcinoma (LIHC) cohort, robustly reproducing previous findings reported by GISTIC2.0 (Ally et al. 2017). The overall pattern of recurrent broad and focal CNAs described in the TCGA study was similar to earlier reports, confirming the specific copy number profile for hepatocellular carcinoma (HCC) (Chiang et al. 2008; Guichard et al. 2012; Wang et al. 2013; Totoki et al. 2014; Schulze et al. 2015). By using GISTIC2.0, the most frequent broad alterations in LIHC were gains at 1q (61%) and 8q (52%), and losses at 8p (70%) and 17p (56%) (Supplemental_Table_S1). Recurrent focal amplifications involved the well-characterized driver oncogenes *CCND1* and *FGF19* (11q13.3), *MYC* (8q24.21), *MET* (7q31.2), *VEGFA* (6p21.1) and *MCL1* (1q21.3), and the most recurrent deletions included tumor suppressor genes such as *RB1* (13q14.2) and the *CDKN2A* (9p21.3) genes (Supplemental_Table_S2).

By applying the default parameters of CNApp to the LIHC dataset and selecting chromosome arms as genomic regions to assess broad events, we consistently found copy number gains at 1q (56%) and 8q (46%), and copy number losses at 8p (62%) and 17p (47%) as the most frequent alterations (Figure 3A). The slightly lower rate tendency of broad CNAs from CNApp as compared to GISTIC2.0 also appeared in the subsequent recurrent broad alterations (Supplemental_Table_S1). For instance, GISTIC2.0 significantly detected gains with rates between 25-40% on eight additional chromosome-arms, including 5p, 5q, 6p, 20p, 20q, 7p, 7q, and 17q, which were identified by CNApp in 20-30% of the samples. Similarly, GISTIC2.0 significantly detected broad deletions at frequencies between 20-40% on 18 additional chromosome-arms, of which 4q, 6q, 9p, 13q, 16p, and 16q losses were observed at ≥20% by CNApp, and the rest of them displayed rates between 10-20%. In this case, discrepancies in CNA frequencies were expected considering the lower copy number amplitude thresholds used by GISTIC2.0 in comparison with the CNApp default cutoffs (|0.1| vs |0.2|, corresponding to ~2.14/1.8 copies vs 2.3/1.7 copies, respectively). Indeed, previous reports analyzing CNAs in other HCC cohorts and using greater copy number thresholds, showed frequencies of alterations similar to those estimated by CNApp (Chiang et al. 2008; Guichard et al. 2012; Wang et al. 2013; Schulze et al. 2015). To assess the impact of modifying CNApp amplitude thresholds, we next re-run the software dropping the minimum copy number values to |0.1|. As expected, the overall number of broad alterations increased, reaching frequency values similar or even higher than those reported by GISTIC2.0 (Figure 3B and Supplemental_Table_S1). Of note, such drop from 0.2 to 0.1 might facilitate the identification of subclonal genomic imbalances, which are very frequent among tumor samples (McGranahan and Swanton 2017), though it can also increase the number of false positive calls. Furthermore, we assessed whether the identification of broad events was affected by two additional parameters: (i) the relative length to classify a segment as *arm-level* alteration, and (ii) the re-segmentation provided by CNApp. As expected, increasing the percentage of chromosome arm required to classify a CNA segment as *arm-level* (from ≥ 50% to ≥ 70%) or skipping the re-segmentation step led to an underestimation of some broad events, whereas decreasing the percentage of chromosome arm (from ≥50% to ≥40%) resulted in the opposite (Supplemental_Fig_S3A-C and Supplemental_Table_S1).

As far as focal CNAs are concerned, CNApp and GISTIC2.0 use different strategies to quantify their recurrence. Therefore, the comparison between the two methods was evaluated in a more indirect manner. GISTIC2.0 constructs minimal common regions (also known as 'peaks') that are likely to be altered at high frequencies in the cohort, which are scored using a Q-value and may present a wide variety of genomic lengths (Mermel et al. 2011). Instead, CNApp allows dividing the genome in windows of different sizes, calculating an average of the copy number amplitudes of segments included within the selected windows. We reasoned that considering the length of GISTIC2.0 reported 'peaks', CNApp might also be capable to identify focal recurrently altered regions by dividing the genome in windows of a relatively small size. To test our hypothesis, we asked CNApp to calculate the frequency of focal gains and losses by dividing the genome by sub-cytobands. As a result, CNApp consistently localized the most frequently altered sub-cytobands (found in 10-25% of samples), including gains at 1q21.3 (25%), 8q24.21 (17%, *MYC*), 5p15.33 (13%, *TERT*), 11q13.3 (12%, *CCND1*/*FGF19*) and 6p21.1 (11%, *VEGFA*), and losses at 13q14.2 (20%, *RB1*), 1p36.11 (18%, *ARID1A*), 4q35.1 (17%, *IRF2*) and 9p21.3 (14%, *CDKN2A*), which are in agreement with previous studies in HCC (Figure 3C and Supplemental_Table_S2) (Chiang et al. 2008; Guichard et al. 2012; Wang et al. 2013; Schulze et al. 2015).

Compared to GISTIC2.0, CNApp reported 14 of the 27 significant amplifications and 14 of the 34 significant deletions at rates >10%, and the remaining alterations displaying rates between 4-10% (Supplemental_Table_S3) (Wang et al. 2013). Most importantly, regions with the highest frequency detected by CNApp showed a good match with lowest GISTIC2.0 Q-residual values, indicating that the most significant 'peaks' identified by GISTIC2.0 were actually included in the most recurrently altered sub-cytobands reported by CNApp.

As previously suggested, recurrent focal alterations often occur at lower frequencies than broad events (Beroukhim et al. 2010). However, previous studies describing the genomic landscape of HCC mostly focused on high-level focal CNAs (from >3 copies for gains and from <1.3 copies for losses), thus reporting lower frequencies than those estimated by CNApp (Chiang et al. 2008; Guichard et al. 2012; Schulze et al. 2015). Interestingly, excluding the low-level alterations and evaluating only the moderate and high-amplitude events (≥3 and ≤1 copies), frequencies dropped to values closer to those previously reported (Figure 3D and Supplemental_Table_S2). Amplifications reached maximum rates of 11%, whereas losses ended up at rates of ~2%, in consistence with the observation that high-level CNAs are relatively rare (Zack et al. 2013). Top recurrent gains involved sub-cytobands 1q21.3 (11%) and 8q24.21 (11%, *MYC*), 11q13.3 (7%, *CCND1*/*FGF19*), and 5p15.33 (5%, *TERT*). Recurrent losses estimated at ~2% of the samples included 13q14.2 (*RB1*), 9p21.3 (*CDKN2A*), 4q35.1 (*IRF2*), and 8p23.1. Slight discrepancies between frequencies might be explained by minimal variability in the copy number threshold.

Classification of colon cancer according to CNA scores and genomic regions

A proposed taxonomy of colorectal cancer (CRC) includes four consensus molecular subtypes (CMS), mainly based on differences in gene expression signatures. Accordingly, each CMS shows specific molecular features such as microsatellite instability (MSI) status, CpG island methylator phenotype (CIMP) levels, somatic CNAs and non-synonymous mutations. Briefly, CMS1 includes the majority of hypermutated tumors showing MSI, high CIMP, and low levels of CNAs; CMS2 and 4 typically comprise microsatellite stable (MSS) tumors with high levels of CNAs; and finally, mixed MSI status and low levels of CNAs and CIMP are associated with CMS3 tumors (Guinney et al. 2015). Using a representative cohort of 309 colon cancers from the TCGA Colon Adenocarcinoma (COAD) cohort (Cancer and Atlas 2012) with known CMS classification (CMS1, N = 64; CMS2 N = 112; CMS3 N = 51; CMS4 N = 82) and MSI status, we asked CNApp to generate a genome-wide frequency plot after re-segmentation using the default copy number thresholds and excluding segments smaller than 500 Kbp to avoid technical background noise. CNA profiles were generated using genomic regions defined by chromosome arms. As expected, the frequency plot displayed the most commonly altered genomic regions in sporadic CRC (Camps et al. 2008; Cancer and Atlas 2012; Ried et al. 1996; Meijer et al. 1998; Nakao et al. 2004). By assessing the broad CNA events in the entire cohort, we observed that the most frequently altered chromosome arms were gains of 7p, 7q, 8q, 13q, 20p, and 20q, and losses of 8p, 17p, 18p, and 18q, occurring in more than 30% of the samples (Figure 4A). Focal CNAs were obtained by generating genomic regions by sub-cytobands. Of note, five out of six genomic losses and five out of 18 genomic gains contained deletions and amplifications, respectively, identified by GISTIC2.0 in the COAD TCGA cohort.

Subsequently, we performed integrative analysis of genomic imbalances, CMS groups, and CNA scores. By using CNApp, we assessed whether CNA scores were able to classify colon cancer samples according to their CMS. While BCS established significant differences between CMS paired comparisons ($P \leq 0.0001$, Student's t-test), FCS poorly discern CMS1 from 3 and CMS2 from 4 (Figure 4B and Supplemental_Fig_S4A). Thus, we reasoned that broad CNAs rather than focal were able to better discriminate between different CMS groups. In fact, the distribution of CMS groups based on BCS resembled the distribution of somatic CNA counts defined by GISTIC2.0 (Guinney et al. 2015), which agrees with the observation that BCS highly correlates with the fraction of altered genome (Supplemental_Fig_S1A). Subsequently, we integrated the BCS and the CMS groups with the microsatellite status. Our results showed an average BCS of 1.51□2.11 and 10.25□5.92 for MSI (N = 72) and MSS (N = 225) tumors, respectively. In addition, a BCS of 4, corresponding to the 90th percentile in the MSI sample set, was able to differentiate MSI and MSS tumors. Applying this cutoff, 186 out of 225 (83%) of MSS tumors showed a BCS greater than 4 (Figure 4C). In contrast, 39 (17%) MSS tumors showed a BCS value of 4 or lower, corresponding to three CMS1, six CMS2, 18 CMS3 and 12 CMS4 tumors, further demonstrating the existence of MSS tumors with a very low CNA burden. When we assessed the level of focal alterations in this subset of MSS samples by considering the 90th percentile of FCS in the MSI group (37.2), we could determine that eight of these MSS tumors showed high FCS, thus reducing the percentage of MSS tumors with overall low copy number changes to 13%. On the other hand, seven MSI tumors showed BCS higher than 4. Among these, five samples displayed genomic imbalances typically associated with the CRC canonical pathway, including a focal amplification of *MYC*, unveiling tumors with co-occurrence of MSI and extensive genomic alterations (Trautmann et al. 2006). Our dataset comprised nine out of 51 CMS3 tumors with MSI. Intriguingly, two of them showed focal deletions on chromosome 2 involving *MSH2* and *MSH6*, suggesting the inactivation of these mismatch repair genes through a focal genomic imbalance. In fact, 46% of CMS3 MSS tumors showed BCS below 4, in agreement with the finding that CMS3 tumors display low levels of somatic CNAs.

CNApp enable the identification of possible sample misclassifications by integrating CMS annotation and *BRAF*-mutated sample status.. As expected, CMS1 cases were enriched for *BRAF* mutation. Nevertheless, two CMS4 samples also showed mutations in *BRAF*. One of these samples showed a BCS of 11, displaying canonical CNAs. In contrast, the other CMS4 BRAF-mutated sample showed MSI and a BCS of 0, similar features as in CMS1. Likewise, four *BRAF* WT samples, classified within the CMS4 group, displayed MSI and a BCS of 0, thus being candidates to be labeled as CMS1 based on the levels of CNAs (Figure 4D). These disparities are of utmost importance since recent studies reported that high copy number alterations correlate with reduced response to immunotherapy (Davoli et al. 2017). Importantly, it has been suggested that MSI status might be predictive of positive immune checkpoint blockade response in advanced CRC, probably due to the low levels of CNA usually presented by MSI tumors (Le et al. 2015). We next asked CNApp to compare genomic regions differentially represented in the four CMS groups based on a Student's t-test or Fisher's test with adjusted p-value. By applying a Student's t-test, we could observe that CMS1 resembled CMS3, except for the gain of chromosome 7 and the loss of 18q, which were the alterations that commonly appeared in CMS3 samples with BCS above 4 ($P \leq 0.001$, Student's t-test) (Supplemental_Fig_S4B). Even though only subtle CNA differences between CMS2 and CMS4 were identified, the loss of 14q was significantly more detected in CMS2 (42%) than in CMS4 (17.1%) ($P \leq 0.005$, Student's t-test) (Supplemental_Fig_S4B). Visually exploring the heatmap plot and further analyzing specific regions, we observed that the gain of 12q was more frequently associated with CMS1 than CMS2 ($P \leq 0.005$, Student's t-test), in agreement with previous studies reporting that the gain of chromosome 12 is associated with microsatellite unstable tumors (Supplemental_Fig_S4B) (Trautmann et al. 2006). Intriguingly, the gain of the chromosome arm 20q alone mimicked the distribution of somatic CNAs defined by GISTIC2.0 across consensus subtype samples

(Figure 4E) (Guinney et al. 2015). In fact, chromosome arm 20q was gained in 99.1%, 70.7%, 39.2%, and 10.9% of CMS2, CMS4, CMS3 and CMS1 tumors, respectively.

Finally, we applied machine learning-based prediction models to classify samples by their MSI status or CMS. BCS predicted MSI status with a global accuracy of 82.2%. This was consistent with the fact that BCS was able to distinguish CMS1 from CMS2 with 89.2% of accuracy. However, when we tested the performance of BCS to predict any CMS group, the accuracy was only 47.5%, indicating that BCS alone is a poor predictive variable to assess CMS. We then used the most discriminative descriptive regions among CMS groups (i.e., 13q, 17p, 18, and 20q), and reached an accuracy to correctly predict CMS of 55%. In fact, the occurrence of these genomic alterations was able to differentiate CMS2 from CMS4 with an accuracy of 70%, and CMS1 from CMS3 with a 72.3% accuracy. As expected, this set of genomic alterations distinguished CMS1 from CMS2 samples with an accuracy of 95%. Altogether, these data suggest that CNApp might provide insight into further classifying CRC samples in CMS groups.

DISCUSSION

Here we present CNApp, a web-based computational approach to analyze and integrate CNAs associated with molecular and clinical variables. CNApp calculates CNA scores to quantify focal, broad and global levels of alterations for each individual sample after an optional process of re-segmentation. Moreover, CNApp utilizes genomic imbalances selected by the user to assess classifier variables by computing machine learning-based models. Although CNApp has been developed using segmented genomic copy number data obtained from SNP-arrays, the software is also able to accommodate segmented data from next-generation sequencing.

Overall, CNApp was benchmarked by analyzing a pan-cancer TCGA dataset with more than 10,000 samples, being able to cluster major tumor types according to CNA patterns.

Moreover, our results demonstrate the reliability of CNApp in identifying regions encompassing the most recurrent CNAs. The software successfully reproduced the well-characterized genomic profile of HCC and CRC, considering both broad and focal events. Although CNApp has not been developed to define the precise boundaries of focal events, the software is capable to detect which regions are likely to contain the most recurrent alterations. However, we acknowledge that the characterization of focal alterations potentially containing driver events performed by GISTIC2.0 is more accurate than the genomic windows provided by CNApp. Thus, despite the in-depth comparison described here, we consider CNApp as a complementary tool rather than a replacement for GISTIC2.0.

Finally, applying CNApp to a colon cancer dataset for which clinical features were known allowed the determination of a BCS value of 4 to potentially discriminate MSI from MSS tumors. Most importantly, due to the inverse correlation between MSI and aneuploidy in CRC, our results suggest that this BCS value could be established as a cutoff to define the edge between low and high aneuploid tumors. Nevertheless, these results ought to be further validated in an independent cohort. Since high levels of aneuploidy correlate with immune evasion markers, quantification of CNAs and their association with molecular and clinical features might be of extreme relevance. In fact, specific genomic regions defined by CNApp contributed to classify the consensus molecular subtypes. This is of clinical interest as it is known that CMS1 microsatellite unstable tumors might show a positive response to immuno-related treatments. Therefore, we believe that CNApp enables not only the fundamental analysis of CNA profiles, but also the functional understanding of CNAs in the context of clinical samples and their potential use as biomarkers.

METHODS

Data set availability

CNA data from TCGA: pan-cancer cohort

Affymetrix SNP6.0 array copy number segmented data (Level 3) from 10,635 samples spanning 33 cancer types from TCGA pan-cancer dataset were downloaded from Genomic Data Commons (National Cancer Institute, NIH) (Grossman et al. 2016). This dataset included the 370 Liver Cancer-Hepatocellular Carcinoma (LIHC) samples used for the analysis of recurrent CNAs and the subset of 309 samples from Colon Adenocarcinoma (COAD) for which the colorectal cancer consensus molecular subtype (CMS) was known (Guinney et al. 2015).

GISTIC data from TCGA: LIHC cohort

GISTIC 2.0.22 (Ally et al. 2017) copy number results (Level 4) of the 370 LIHC samples, were downloaded from the Broad Institute GDAC Firehose. Parameters used for the analysis are detailed in the same GDAC repository. Specifically, parameters conditioning the definition of the CNAs and of interest for our comparison were publicly reported with the following values: *amplification* and *deletion thresholds*: 0.1; *broad length cutoff*: 0.7; *joint segment size*: 4.

Software and tool availability

CNApp can be accessed at http://bioinfo.ciberehd.org/CNApp. It was developed using Shiny R package (version 1.1.0), from R-Studio (Chang et al. 2018). The tool was applied and benchmarked while using R version 3.4.2 (2017-09-28) -- "Short Summer". List of packages, libraries and base coded are freely available at GitHub, and instructions for local installation are also specified.

CNA scores computation

Segments resulted from re-segmentation (or original segments from input file when re-segmentation is skipped) are classified in *chromosomal*, *arm-level* and *focal* events by considering the relative length of each segment to the whole-chromosome or chromosome arm. Using default parameters, segments are tagged as *chromosomal* when 90% or more of the chromosome is affected; as *arm-level* when 50% or more of the chromosome arm affected; and as *focal* when affecting less than 50% of the chromosome arm. Percentages for relative lengths are customizable. Broad (chromosomal and arm-level) and focal alterations are then weighted according to their amplitude values (*seg.mean*) and taking into account copy number amplitude ranges defined by CNA calling thresholds and specified in Supplemental_Methods.

Broad CNA Score (BCS): for a total *N* of broad events in a sample (*x*), it equals to the summation of segments weights (*A*) in that corresponding sample and being *i* the corresponding segment:

$$BCS(x) = \sum_{i=1}^{N} A_i$$

Focal CNA Score (FCS): same as in BCS, with an additional pondering value *L* included to the summation, which captures the relative size of the chromosome-arm coverage of each focal CNA (according to weights specified in Supplemental_Methods):

$$FCS(x) = \sum_{i=1}^{N} A_i \cdot L_i$$

Global CNA Score (GCS): for a sample *x*, it is calculated as the summation of normalized BCS and FCS values, where *meanBCS* and *meanFCS* stand for mean values of BCS and FCS from total samples, respectively, and *sdBCS* and *sdFCS* stand for standard deviation values of BCS and FCS from total samples, respectively:

$$normBCS(x) = \frac{BCS(x) - meanBCS}{sdBCS} \qquad normFCS(x) = \frac{FCS(x) - meanFCS}{sdFCS}$$

$$GCS(x) = \sum_{i=1}^{N} normBCS_i + normFCS_i$$

Genomic windows computation

Region profiling section allows genome segmentation analysis by user-selected windows (i.e. arms, half-arms, cytobands, sub-cytobands, and 40Mb till 1Mb). In order to do that, windows files were generated for each option and genome build (*hg19* and *hg38*). Cytobands file *cytoBand.txt* from UCSC page and for both genome builds was used as mold to compute regions (Casper et al. 2017).

Segmented samples are transformed into genome region profiles using genomic windows selected by user. Segments from each sample are consulted to assess whether or not overlap with the window region. Thus, window-means (*W*) are computed for each genomic window by collecting segments (*t*) overlapping with window-region (*i*). Segments with *loc.start* or *loc.end* position falling within the region are collected, as well as those segments embedding the entire region. At this point, the summation of each segment-mean (*S*) corrected by the relative window-length (*L*) affected by the segment length (*l*) is performed:

$$W(i) = \sum_{t=1}^{n} S_t \cdot \frac{l_t}{L(i)}$$

Descriptive regions assessment

Potential descriptive regions between groups defined by the annotated variables provided in the input file can be studied and *P*-values are presented to evaluate significance in differentially altered regions between those groups. The alterations can be considered as (1) numerical continuous (*seg.mean* values) and (2) categorical variables (gains, losses and non-altered). In the first case, to assess statistical significance between groups

Student's T-test is applied, whereas in the second situation the significance is assessed by applying the Fisher's exact test. False discovery rate (FDR) adjustment is performed using the Benjamini-Hochberg (BH) procedure in both cases and corrected *P*-values (*Adj.p-value*) or non-corrected *P*-values (*p-values*) are displayed by user selection.

Machine learning-based classifier models

We used the *randomForest* R package (Liaw and Wiener 2002) to compute machine learning classifier models. Variables to define sample groups must be selected, as well as at least one classifier variable. Model construction is performed 50-times and training set is changed by iteration. In order to compute model and select training set, multiple steps and conditions have to be accomplished:

i. total *N* samples divided by *G* groups depicted by group-defining variable must be higher than n samples from the smaller group:

$$P = \frac{N}{G} \; ; \; P > n$$

ii. If condition above is not accomplished, then *P* is set to 75% of n:

$$\text{if } P \le n \text{ then } P = n \cdot 0.75$$

iii. *P* term must be higher than one, and *N* must be equal or higher than 20:

$$P > 1 \text{ or } N \ge 20$$

iv. Classifier variables, when categorical, shall not have higher number of tags (*Z*) than groups defined (*G*) by group-defining variable:

$$Z < G$$

v. Training set (*T*) is computed and merged for each group (*g*) from groups (*G*) defined by group variable, extracting *P* samples from *g* as follows:

$$t\ (g) = P \text{ samples from } g \qquad\qquad T = \sum_{i=1}^{g} t_i$$

After model computation, contingency matrix with prediction and reference values by group is created to compute accuracy, specificity and sensitivity by group.

Ethics approval and consent to participate

Ethics approval was not required for this study.

Consent for publication

Not applicable.

Conflicts of interests

Dr. Llovet is receiving research support from Bayer HealthCare Pharmaceuticals, Eisai Inc, Bristol-Myers Squibb and Ipsen, and consulting fees from Eli Lilly, Bayer HealthCare Pharmaceuticals, Bristol-Myers Squibb, EISAI Inc, Celsion Corporation, Exelixis, Merck, Ipsen, Glycotest, Navigant, Leerink Swann LLC, Midatech Ltd, and Nucleix.

Funding

This work has been supported by grants from the European Commission (PCIG11-GA-2012-321937), the Instituto de Salud Carlos III and co-funded by the European Regional Development Fund (ERDF) (CP13/00160, PI14/00783, PI17/01304, PI17/00878), the CIBEREHD program, the CERCA Program (Generalitat de Catalunya), the Agència de

Gestió d'Ajuts Universitaris i de Recerca, Generalitat de Catalunya (2017 SGR 1035, 2017 SGR 21, and 2017 SGR 653), PERIS (SLT002/16/00398, Generalitat de Catalunya), Fundación Científica de la Asociación Española Contra el Cáncer (GCB13131592CAST). SF-E was supported by a contract from CIBEREHD. LB and MD-G were supported by the Agència de Gestió d'Ajuts Universitaris i de Recerca - AGAUR- from Generalitat de Catalunya (2016BP00161 and 2018FI B1_00213, respectively). RE-F is supported by the Spanish National Health Institute (FPI grant BES-2017-081286). CIBEREHD is funded by the Instituto de Salud Carlos III. This article is based upon work from COST Action CA17118, supported by COST (European Cooperation in Science and Technology). JML is supported by the European Commission (EC)/Horizon 2020 Program (HEPCAR, Ref. 667273-2), U.S. Department of Defense (CA150272P3), National Cancer Institute (P30-CA196521), Samuel Waxman Cancer Research Foundation, Spanish National Health Institute (SAF2016-76390) and Generalitat de Catalunya/AGAUR (SGR-1162 and SGR-1358).

Acknowledgements

The authors would like to thank Dr. Rodrigo Dienstmann from Vall d'Hebron Institute of Oncology, Barcelona, Spain for providing the CMS and clinical information for the subset of samples included in the COAD cohort from TCGA. The work was carried out at the Esther Koplowitz Centre, Barcelona.

Authors contributions

SF-E, LB and JC designed the study and analyzed the data. SF-E, LB, MV-C and MD-G designed, generated and implemented the package, and analyzed the data. EH-I and RE-F tested the software. JJL supervised the software implementation. AC, and JMLl critically reviewed the software implementation and the data output. SF-E, LB, SC-B and JC wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, Holt RA, Jones SJM, Lee D, Ma Y, et al. 2017. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**: 1327–1341.e23.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, et al. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**: 899–905.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, Barillot E. 2012. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**: 423–425.

Buccitelli C, Salgueiro L, Rowald K, Sotillo R, Mardin BR, Korbel JO. 2017. Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. *Genome Res* **27**: 501–511.

Cairncross G, Wang M, Shaw E, Jenkins R, Brachman D, Buckner J, Fink K, Souhami L, Laperriere N, Curran W, et al. 2013. Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: Long-term results of RTOG 9402. *J Clin Oncol* **31**: 337–343.

Camps J, Grade M, Nguyen QT, Hormann P, Becker S, Hummon AB, Rodriguez V, Chandrasekharappa S, Chen Y, Difilippantonio MJ, et al. 2008. Chromosomal Breakpoints in Primary Colon Cancer Cluster at Sites of Structural Variants in the Genome. *Cancer Res* **68**: 1284–1295.

Cancer T, Atlas G. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**: 330–7.

Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, et al. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**: 413–421.

Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, et al. 2017. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* **46**: D762–D769.

Chang W, Cheng J, Allaire J, Xie Y, McPherson J. 2018. shiny: Web Application Framework for R. R package version 1.1.0.

Chiang DY, Villanueva A, Hoshida Y, Peix J, Newell P, Minguez B, LeBlanc AC, Donovan DJ, Thung SN, Sole M, et al. 2008. Focal Gains of VEGFA and Molecular Classification of Hepatocellular Carcinoma. *Cancer Res* **68**: 6779–6788.

Davoli T, Uno H, Wooten EC, Elledge SJ. 2017. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science (80-)* **355**: eaaf8399.

Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. 2016. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**: 1109–1112.

Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad I Ben, Calderaro J, Bioulac-Sage P, Letexier M, Degos F, et al. 2012. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet* **44**: 694–698.

Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al. 2015. The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**: 1350–1356.

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. 2018. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**: 291–304.e6.

Krijgsman O, Carvalho B, Meijer GA, Steenbergen RDM, Ylstra B. 2014. Focal chromosomal copy number aberrations in cancer-Needles in a genome haystack. *Biochim Biophys Acta - Mol Cell Res* **1843**: 2698–2704.

Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, et al. 2015. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* **372**: 2509–2520.

Liaw A, Wiener M. 2002. Classification and Regression by randomForest. *R News* **2**: 18–22.

McGranahan N, Swanton C. 2017. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**: 613–628.

Meijer GA, Hermsen MA, Baak JP, van Diest PJ, Meuwissen SG, Belien JA, Hoovers JM, Joenje H, Snijders PJ, Walboomers JM, et al. 1998. Progression from colorectal adenoma to carcinoma is associated with non- random chromosomal gains as detected by comparative genomic hybridisation. *J Clin Pathol* **51**: 901–909.

Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**: R41.

Nakao K, Mehta KR, Fridlyand J, Moore DH, Jain AN, Lafuente A, Wiencke JW, Terdiman JP, Waldman FM. 2004. High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*.

Olshen AB, Venkatraman ES, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557–572.

Popova T, Manié E, Stoppa-Lyonnet D, Rigaill G, Barillot E, Stern MH. 2009. Genome Alteration Print (GAP): A tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol* **10**: R128.

Ried T, Hu Y, Difilippantonio MJ, Ghadimi BM, Grade M, Camps J. 2012. The consequences of chromosomal aneuploidy on the transcriptome of cancer cells. *Biochim Biophys Acta* **1819**: 784–93.

Ried T, Knutzen R, Steinbeck R, Blegen H, Schröck E, Heselmeyer K, du Manoir S, Auer G. 1996. Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes, Chromosom Cancer* **15**: 234–245.

Sansregret L, Vanhaesebroeck B, Swanton C. 2018. Determinants and clinical implications of chromosomal instability in cancer. *Nat Rev Clin Oncol* **15**: 139–150.

Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27**: 2648–2654.

Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, Couchy G, Meiller C, Shinde J, Soysouvanh F, et al. 2015. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet* **47**: 505–511.

Stratton MR, Campbell PJ, Futreal PA, Andrew F P. 2009. The cancer genome. *Nature* **458**: 719–724.

Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, et al. 2018a. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**: 676–689.e3.

Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, et al. 2018b. Genomic and Functional Approaches to Understanding Cancer

Aneuploidy. *Cancer Cell* **33**: 676–689.e3.

Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, Tsuji S, Donehower LA, Slagle BL, Nakamura H, et al. 2014. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat Genet* **46**: 1267–1273.

Trautmann K, Terdiman JP, French AJ, Roydasgupta R, Sein N, Kakar S, Fridlyand J, Snijders AM, Albertson DG, Thibodeau SN, et al. 2006. Chromosomal instability in microsatellite-unstable and stable colon cancer. *Clin Cancer Res* **12**: 6379–6385.

van de Wie MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B. 2007. CGHcall: Calling aberrations for array CGH tumor profiles. *Bioinformatics* **23**: 892–894.

Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, et al. 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci* **107**: 16910–16915.

Wang K, Lim HY, Shi S, Lee J, Deng S, Xie T, Zhu Z, Wang Y, Pocalyko D, Yang WJ, et al. 2013. Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. *Hepatology* **58**: 706–717.

Zack TTI, Schumacher SES, Carter SLS, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang C-Z, Wala J, Mermel CH, et al. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**: 1134–1140.

Zhang Z, Hao K. 2015. SAAS-CNV: A Joint Segmentation Approach on Aggregated and Allele Specific Signals for the Identification of Somatic Copy Number Alterations with Next-Generation Sequencing Data ed. E. Wang. *PLoS Comput Biol* **11**: e1004618.

FIGURE LEGENDS

Figure 1: CNApp workflow. The diagram depicts the overall processes performed by CNApp and indicates the output for each section.

Figure 2: Analysis of the TCGA pan-cancer dataset and clustering by tumor type. CNApp outputs to characterize pan-cancer 10,635 samples including 33 TCGA cancer types. **A)** Broad, Focal and Global CNA scores (BCS, FCS and GCS, respectively) distribution across the 33 cancer types. **B)** Genome-wide chromosome arm CNA profile heatmap for 10,635 samples considering broad and focal events. Annotation tracks for FCS, BCS and GCS are presented. **C)** Arm regions frequencies as percentages relative to the TCGA pan-cancer dataset (red for gains and blue for losses). **D)** Heatmap plot showing 20 out of the 33 TCGA cancer type profile correlations, by Pearson's method, hierarchically clustered by tumor type. Gastrointestinal, gynecological and squamous types are clustering consistently in their respective groups.

Figure 3: Identification of recurrent broad and focal CNAs. Calculation of broad and focal CNA frequencies using several parameters in CNApp in order to describe the genomic landscape of LIHC. **A)** CNApp frequencies for chromosome arm regions using default cutoffs, corresponding to 2.3/1.7 copies for gains and losses, respectively. **B)** CNApp frequencies for chromosome arm regions relaxing cutoffs to make them equivalent to those of GISTIC2.0. **C)** CNApp frequencies of focal events using default thresholds and sub-cytobands genomic regions. **D)** Frequencies of focal events from moderate- to high-amplitude levels using sub-cytobands genomic regions.

Figure 4: Genomic characterization of colon cancer according to the CMS

classification. A) Arm-region frequencies of 309 colon cancer samples using CNApp default thresholds for CNAs. **B)** BCS distribution by CMS sample groups. Significance is shown as p-value ≤ 0.001 (***); p-value ≤ 0.01 (**); p-value ≤ 0.05 (*); p-value > 0.05 (ns). **C)** Number of gained and lost chromosome arms for each sample distributed according to the BCS values. Note that a cutoff at 4 is indicated with a black line. Annotation tracks for microsatellite instability (msi), *BRAF* mutated samples (braf_mut), CMS groups (cms_label), FCS and BCS are displayed. **D)** Genome-wide profiling by chromosome arms distributed according to the CMS group. Annotation tracks for microsatellite instability (msi), *BRAF* mutated samples (braf_mut), CMS groups (cms_label), FCS and BCS are displayed. Sample-to-sample correlation heatmap plot by Pearson's method is shown below. **E)** Distribution of CNA values affecting 20q according to the CMS groups. Significance is shown as p-value ≤ 0.001 (***); p-value ≤ 0.01 (**); p-value ≤ 0.05 (*); p-value > 0.05 (ns).

Figure 1

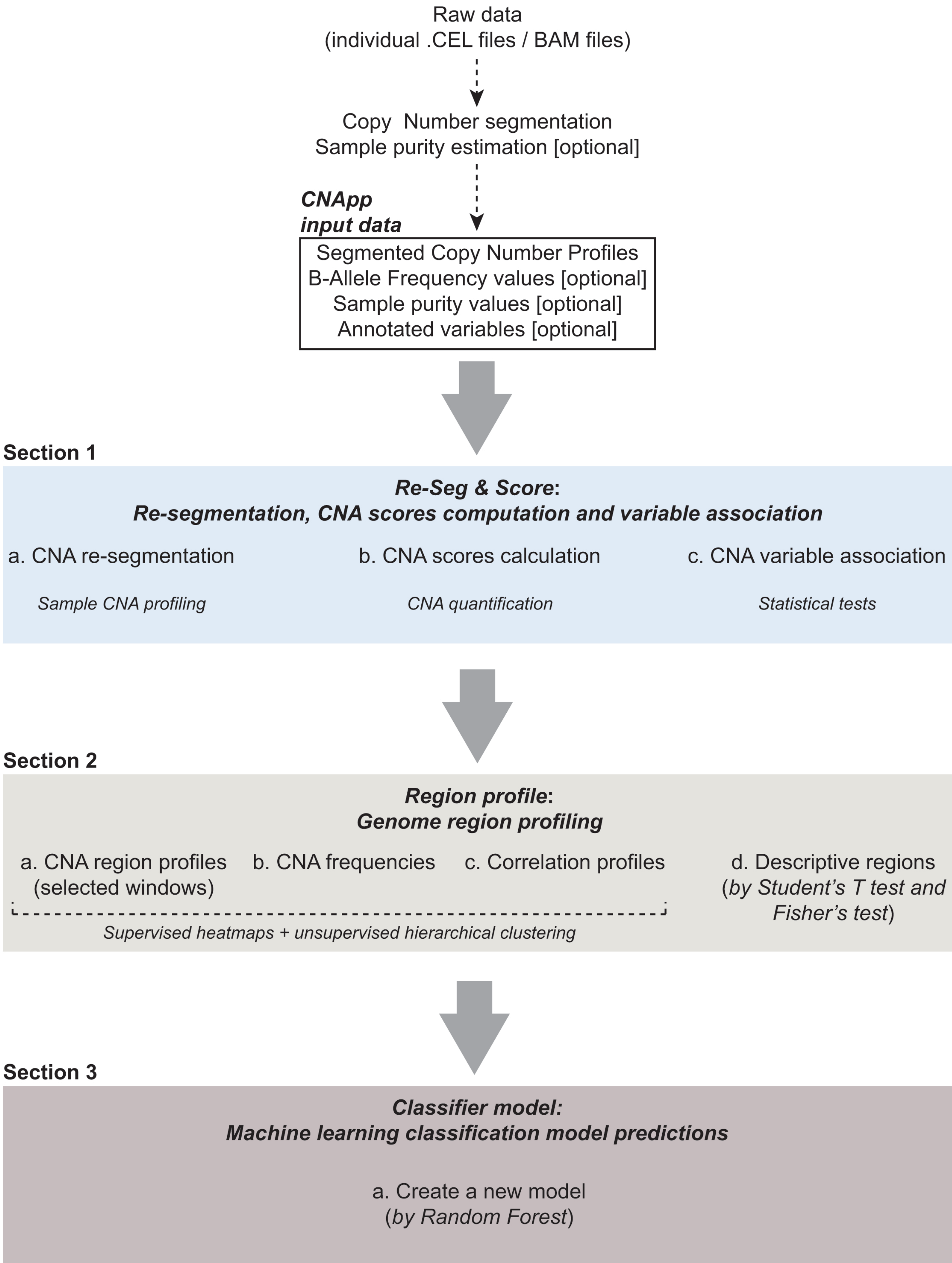

Raw data
(individual .CEL files / BAM files)
Copy Number segmentation
Sample purity estimation [optional]
CNApp
input data
Segmented Copy Number Profiles
B-Allele Frequency values [optional]
Sample purity values [optional]
Annotated variables [optional]
Section 1
Re-Seg & Score:
Re-segmentation, CNA scores computation and variable association
a. CNA re-segmentation
Sample CNA profiling
b. CNA scores calculation
CNA quantification
c. CNA variable association
Statistical tests
Section 2
Region profile:
Genome region profiling
a. CNA region profiles
(selected windows)
b. CNA frequencies
c. Correlation profiles
Supervised heatmaps + unsupervised hierarchical clustering
d. Descriptive regions
(*by Student's T test and*
Fisher's test)
Section 3
Classifier model:
Machine learning classification model predictions
a. Create a new model
(*by Random Forest*)

Figure 2

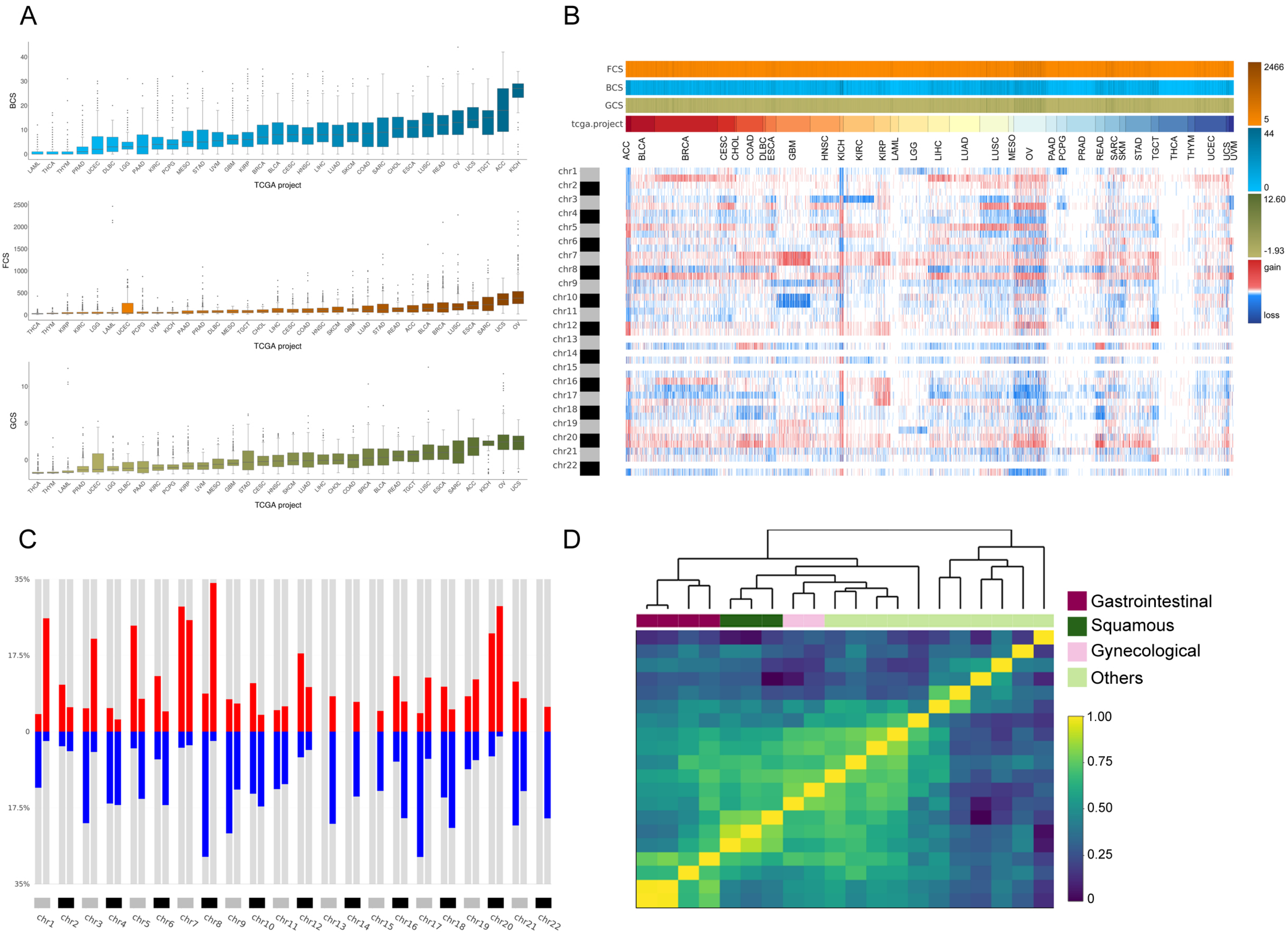
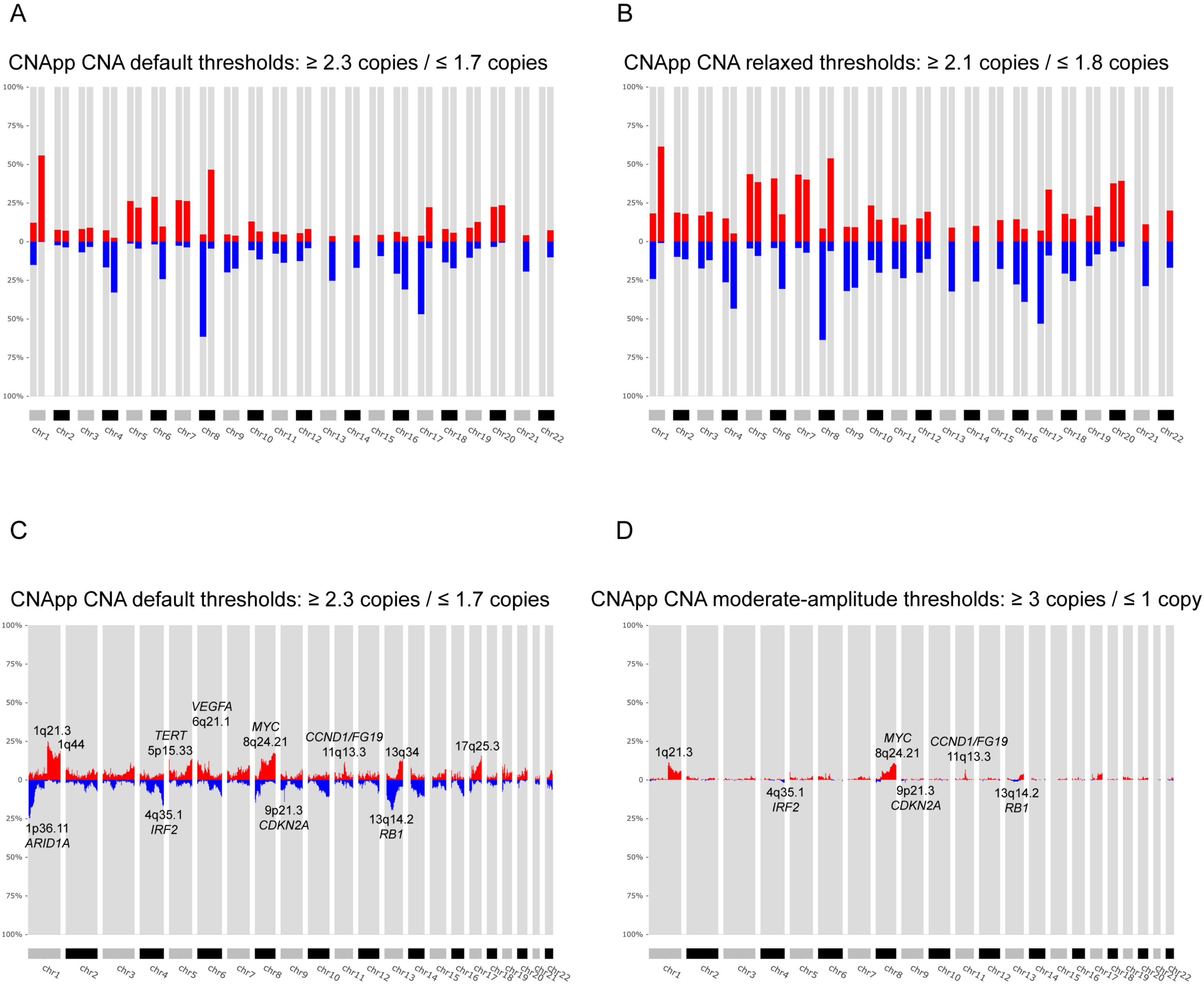

A
BCS
40
30
20
10
0
LAML THCA THYM PRAD UCEC DLBC LGG PAAD KIRC PCPG MESO STAD UVM GBM KIRP BRCA BLCA CESC HNSC LIHC LUAD SKCM COAD SARC CHOL ESCA LUSC READ OV UCS TGCT ACC KICH
TCGA project
FCS
2500
2000
1500
1000
500
0
THCA THYM KIRP KIRC LGG LAML UCEC PCPG UVM KICH PAAD PRAD DLBC MESO TGCT CHOL LIHC CESC COAD HNSC SKCM GBM LUAD STAD READ ACC BLCA BRCA LUSC ESCA SARC UCS OV
TCGA project
GCS
10
5
0
THCA THYM LAML PRAD UCEC LGG DLBC PAAD KIRC PCPG KIRP UVM MESO GBM STAD CESC HNSC SKCM LUAD LIHC CHOL COAD BRCA BLCA READ TGCT LUSC ESCA SARC ACC KICH OV UCS
TCGA project
B
FCS
BCS
GCS
tcga.project
ACC BLCA BRCA CESC CHOL COAD DLBC ESCA GBM HNSC KICH KIRC KIRP LAML LGG LIHC LUAD LUSC MESO OV PAAD PCPG PRAD READ SARC SKM STAD TGCT THCA THYM UCEC UCS UVM
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
2466
5
44
0
12.60
-1.93
gain
loss
C
35%
17.5%
0
17.5%
35%
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
D
Gastrointestinal
Squamous
Gynecological
Others
1.00
0.75
0.50
0.25
0

Figure 3



A

CNApp CNA default thresholds: ≥ 2.3 copies / ≤ 1.7 copies

B

CNApp CNA relaxed thresholds: ≥ 2.1 copies / ≤ 1.8 copies

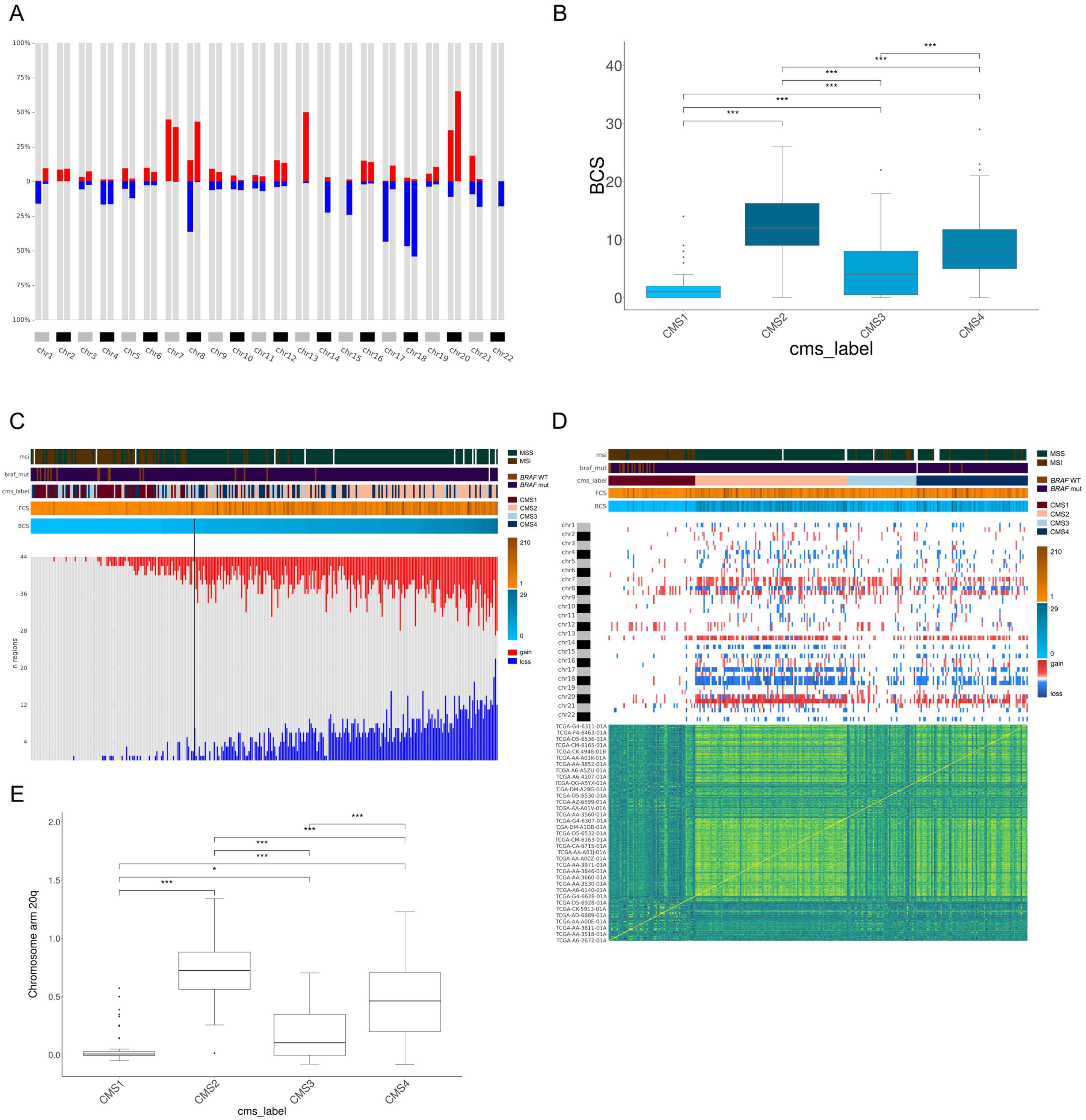
C

CNApp CNA default thresholds: ≥ 2.3 copies / ≤ 1.7 copies

D

CNApp CNA moderate-amplitude thresholds: ≥ 3 copies / ≤ 1 copy

Figure 4


A
100%
75%
50%
25%
0
25%
50%
75%
100%
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
B
BCS
40
30
20
10
0
***
CMS1
CMS2
CMS3
CMS4
cms_label
C
msi
braf_mut
cms_label
FCS
BCS
MSS
MSI
BRAF WT
BRAF mut
CMS1
CMS2
CMS3
CMS4
210
1
29
0
gain
loss
n regions
44
36
28
20
12
4
D
msi
braf_mut
cms_label
FCS
BCS
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22
TCGA-G4-6311-01A
TCGA-F4-6463-01A
TCGA-D5-6536-01A
TCGA-CM-6165-01A
TCGA-CK-4948-01B
TCGA-AA-A01K-01A
TCGA-AA-3852-01A
TCGA-A6-A5ZU-01A
TCGA-A6-4107-01A
TCGA-QG-A5YX-01A
TCGA-DM-A28G-01A
TCGA-D5-6530-01A
TCGA-AZ-6599-01A
TCGA-AA-A01V-01A
TCGA-AA-3560-01A
TCGA-G4-6307-01A
TCGA-DM-A1D8-01A
TCGA-D5-6532-01A
TCGA-CM-6163-01A
TCGA-CA-6715-01A
TCGA-AA-A03J-01A
TCGA-AA-A00Z-01A
TCGA-AA-3971-01A
TCGA-AA-3846-01A
TCGA-AA-3660-01A
TCGA-AA-3530-01A
TCGA-A6-6140-01A
TCGA-G4-6628-01A
TCGA-D5-6928-01A
TCGA-CK-5913-01A
TCGA-AD-6889-01A
TCGA-AA-A00E-01A
TCGA-AA-3811-01A
TCGA-AA-3518-01A
TCGA-A6-2672-01A
MSS
MSI
BRAF WT
BRAF mut
CMS1
CMS2
CMS3
CMS4
210
1
29
0
gain
loss
E
Chromosome arm 20q
2.0
1.5
1.0
0.5
0.0
***
*
CMS1
CMS2
CMS3
CMS4
cms_label